arXiv:2304.01420v1 [cs.CY] 4 Apr 2023

科学家对生成式人工智能在其领域中的潜力的观点

Meredith Ringel
Morris Google Research
merrie@google.com

摘要

生成式人工智能模型，包括大语言模型以及融合文本和其他媒体的多模态模型，即将改变现代生活的诸多方面，包括娱乐、教育、公民生活、艺术以及各种职业。生成式人工智能有望对多个科学学科的发现方法和进程产生实质性影响。我们访谈了来自不同领域（包括物理科学、生命科学和社会科学）的二十位科学家，以了解生成式人工智能技术是否以及如何为他们各自学科的实践增添价值。这不仅包括人工智能可能加速科学发现（即研究）的方式，也包括其职业中的其他方面，例如培养未来学者以及交流科学发现。除了识别生成式人工智能增强科学家当前实践的机会外，我们还请参与者谈谈他们对人工智能的担忧。这些研究发现有助于指导模型和界面的负责任开发，以支持科学教育、探究和交流。

1 引言

过去几年，人工智能技术的能力取得了巨大飞跃，尤其是生成式人工智能系统陆续问世。这些系统能够摄取海量数据并学习统计模式，使模型能够根据自然语言输入（提示词）生成新内容。这些模型可以生成多种类型的内容，包括文本 [8, 41, 51, 14], 图像 [43, 45, 55], 视频 [53, 24], 和音频 [1, 7, 26]。这些人工智能系统日益增强的新能力有望改变现代生活的诸多方面；本文将具体探讨生成式人工智能的进步可能如何改变科学实践，包括生命科学、物理科学和社会科学。

我们访谈了来自不同领域的二十位科学家，以了解生成式人工智能技术是否以及如何为他们各自的学科实践增添价值。这不仅包括生成式人工智能可能加速科学发现（即研究）的方式，也包括他们职业中的其他方面，例如培养未来学者，以及向其他科学家和公众传达科学发现。通过对访谈进行定性分析，我们归纳出一组共同主题，反映了科学家对新型人工智能技术可能如何惠及其职业的看法，以及他们对这类技术可能带来风险的担忧。

我们的参与者预见到生成式人工智能可能带来积极和消极两方面的影响。潜在益处包括新型教育范式；人工智能支持的文献综述，包括跨学科联系与洞见；加速数据集创建和清理的工具；可加快实验和分析各个环节的人工智能辅助编码；新型人工智能辅助数据收集与实验设计方法；以及人工智能增强的写作和同行评审系统。

预印本。正在评审中。

人们担忧的问题包括人工智能辅助作弊及批判性反思机会的丧失、使用人工智能引入偏差、人工智能工具的事实准确性与可信度，以及出版垃圾信息、虚假数据和科学错误信息的泛滥。本研究呈现了在物理科学、生命科学和社会科学各学科实践方面拥有深厚知识的专家观点，所提供的研究发现有助于推动面向科学领域、以人为本且负责任的人工智能工具开发。

在本文中，我们首先讨论生成式人工智能和人工智能辅助科学方面的相关工作。随后介绍参与者的背景以及我们的访谈方法。接着，我们总结研究中涌现的主要主题，涵盖生成式人工智能对科学教育、数据集创建、文献综述、编程、发现和交流的影响。最后，我们讨论科学家如何看待在工作中可能使用的任何人工智能工具所涉及的信任问题，并反思研究发现及其局限性，以及未来的探索方向。

2 相关工作

本研究借鉴以人为本的方法，通过与软件开发人员 [17, 5, 37, 27], 医疗专业人员 [10, 47], 艺术家 [48, 11, 32], 等专家利益相关者合作，了解人工智能对特定终端用户群体的潜力与影响。鉴于本研究特别关注新兴生成式人工智能技术在科学领域的适用性，我们的文献综述聚焦于生成式人工智能的最新进展以及人工智能辅助的科学探究。

2.1 生成式人工智能

变换器神经网络架构的引入 2017 [52] 开启了机器学习的新纪元，使人们能够创建多种强大的预训练生成式人工智能模型（有时也称为基础模型 [6]），这些模型能够以多种模态生成新内容。生成式人工智能最初在语言领域迅速发展（即 GPT 系列语言模型 [8, 41] 及其竞争者 [51, 14]），如今也出现了适用于多种其他模态的模型，包括图像 [43, 45, 55], 、视频 [53, 24], 和音频 [1, 7, 26]。 [1] 这些模型为人机交互和人机协作 [36], 带来了新的机遇，包括增强科学家的能力。

尽管生成式人工智能的能力令人瞩目且在迅速提升（例如，GPT-4 是否可能是迈向通用人工智能的一步 [9], 对此仍有争议），但它仍有诸多局限，包括潜在风险，例如幻觉（捏造内容、不符合事实的输出，如 [18]）、生成有偏差或冒犯性内容、记忆训练数据，以及对训练材料来源的伦理担忧 [3, 54]。在本研究中，我们不仅关注生成式人工智能为提升科学教育、发现和交流带来的机遇，也审慎考量它可能同时带来的挑战与风险。

2.2 人工智能辅助科学

尽管新一代生成式人工智能工具在科学领域的应用尚处于起步阶段（也是本文所述研究的主题），但此前几代人工智能技术，例如基于机器学习的建模、优化、数据分析和早期深度学习技术，已加速了众多科学学科的发现进程，并催生了新的子学科。这些子学科优先采用建模和仿真等计算方法，而非在实验室或野外开展实证研究等传统技术。[2]

人工智能对科学发现产生影响的近期突出案例之一，是 DeepMind 的 AlphaFold 项目 [29], 。该项目利用机器学习，根据氨基酸序列预测蛋白质的三维结构，从而加速了以往需要

[1]请注意，鉴于生成式人工智能近期取得了爆发式进展，我们不打算详尽列举各种模型及其能力，而是从每类技术中选取一些值得关注的实例，以展示最先进技术水平。[2]请注意，人工智能辅助的科学研究数量庞大，难以一一列举；我们将聚焦于少数值得关注且具有代表性的实例，让读者对这一领域有所了解。

耗时在实验室进行蛋白质合成的生物学研究。仅在过去十年间，人工智能就推动了数学 [15], 、天文学 [31,28], 、气候科学 [20], 、保护生物学 [40], 、生物化学 [23], 、医学诊断 [42],、药物发现 [19, 35], 等众多科学学科的新发现与方法发展。

一些科学家已经开始探索生成式人工智能在其领域中的潜力，他们要么利用现成模型，要么针对自身领域对模型进行微调或训练，例如用于天气预测的 ClimaX 模型 [39] 或 GenSLM [56]（一种针对基因组研究进行适配的语言模型）。2022年，谷歌推出了语言模型 Minerva，该模型在多个科学领域的本科水平定量推理问题基准测试中表现出色 [33]。2022年11月，Meta 发布了 Galactica [49], ，这是一种在科学语料库上训练的生成式语言模型，但由于该系统生成了虚假的科学引文和冒犯性的伪科学研究摘要，Meta 仅在数天后就将其撤出公开使用。

人们对生成式人工智能加速科学发展的潜力越来越感兴趣，例如有人提出了诺贝尔图灵挑战这一宏大挑战，旨在“开发一种高度自主的人工智能系统，能够开展顶尖水平的科学研究，其质量应与最优秀的人类科学家所完成的研究难以区分，并且其中一些发现可能达到诺贝尔奖级别，甚至更高” [30]。美国国家科学院最近也考察了现代人工智能对科学发现的影响 [38], ，2023年斯坦福人工智能指数报告 [34], 也对此进行了讨论。该报告专门用一整章反思人工智能促进科学发展方面的最新进展，涉及核聚变 [16], 、矩阵运算 [21], 、芯片设计 [44], 和抗体设计 [46]等领域。

我们的研究以人为本，直接征集并综合物理科学、生命科学和社会科学领域专家的观点，探讨生成式人工智能在其专业领域的教育、发现和交流方面可能带来的益处与风险；这些研究发现可为新型人工智能工具以及人机协作界面和范式的设计提供参考，使其符合科学家（以及社会）的目标。

3 方法

3.1 参与者

我们访谈了二十位科学家（七位女性、十三位男性），他们都在与自身学科相关的领域获得了博士学位。参与者的背景涵盖物理科学（如物理学、地质学、化学）、生命科学（如免疫学、生物信息学、生态学）和社会科学（如社会学、人类学、行为经济学）。访谈特意排除了拥有计算机科学博士学位的参与者，因为开发生成式人工智能技术的人普遍掌握计算机科学实践方面的知识，而本研究旨在获取更广泛的观点。

出于保密原因，所有参与者均隶属于 Alphabet；其中九人是 Alphabet 的全职员工，另有十一人是在大型大学或研究机构任职的终身教职轨或研究型教员，同时在 Alphabet 担任兼职职务（例如，每周一天提供咨询、在固定期限的休假期间全职工作等）。

我们采用滚雪球抽样技术招募参与者，以识别隶属于 Alphabet 的科学家；同时请 Alphabet 的大学关系部门提供一份名单，列出拥有计算机科学以外领域博士学位的访问 faculty（例如兼职顾问、休假访问学者等）。

表1概述了二十位受访者的学科背景和人口统计背景 s.

3.2 访谈结构

一位研究人员在一段为期一个月的时间里完成了全部二十次访谈，时间横跨一月底和2023年二月初。访谈采用半结构式访谈技术，以一套基础问题为提纲，并留出时间，就访谈过程中出现的话题提出后续的自发性深度问题。研究人员提醒参与者，他们可以自由跳过任何问题，也可以在需要时要求澄清。每次访谈持续三十至四十五分钟。

ID	学科	所属机构	性别
P1	行为与实验经济学；市场理论	经济学教授（主要）；Alphabet（兼职）	M
P2	地震科学；地球物理学	地球与行星科学教授（主要）；Alphabet（兼职）	M
P3	免疫学；基因组学	Alphabet（主要）；兼职教授医学系（兼职）	M
P4	材料科学	Alphabet（主要）	F
P5	物理学	Alphabet（主要）	M
P6	水文学	Alphabet（主要）	M
P7	生物化学	生物化学教授（主要）；Alphabet（兼职）	F
P8	电气工程；物理学	航空航天工程教授（主要）；Alphabet（兼职）	F
P9	气候科学	Alphabet（休假）；教授，地球物理学（主要）	M
P10	量子化学	化学教授（主要）；Alphabet（兼职）	M
P11	传播科学；言语语言病理学	传播科学与障碍（主要）；Alphabet（兼职）	M
P12	量子计算硬件	Alphabet（主要）	M
P13	社会学	Alphabet（主要）	M
P14	社会学；科学与技术研究	Alphabet（主要任职）	F
P15	地震学	地球与行星科学教授（主要任职）；Alphabet（兼职）	M
P16	物理学；量子算法	Alphabet（主要任职）	M
P17	医学人类学	Alphabet（主要）	F
P18	干细胞生物学；生物信息学	非营利生物学研究机构（主要）；Alphabet（兼职）	F
P19	物理学；数学；流体动力学	工程与地质科学教授科学（主要）；Alphabet（兼职）	M
P20	地理学	林业研究教授（主要）；Alphabet（兼职）	F

表1：所有访谈参与者都是 Alphabet 的全职或兼职员工；兼职员工主要隶属于大学或研究机构。为保护匿名性，大学名称已删除，但其中包括哈佛大学、斯坦福大学、加州大学伯克利分校和加州理工学院等美国主要研究机构。所有参与者在访谈时均受雇于美国。所有参与者都拥有与其研究领域相关学科的博士学位。

访谈通过视频通话进行，使用 Google Meet。研究人员实时记录访谈笔记，也对访谈过程进行了录音，以防笔记有所遗漏（因研究人员失误，有四场访谈未录音）。参与者没有获得酬谢；他们自愿参与，是因为希望帮助 Alphabet 在人工智能应用于科学职业领域时以负责任的方式进行创新。

访谈的问题主要分为四类。首先，我们收集背景信息（例如人口统计信息、角色）。其次，我们询问参与者所属的科学学科，以进一步了解具体的研究领域和方法论。第三，我们引入生成式人工智能这一主题；在探询参与者对最先进技术及公开可用模型（例如 ChatGPT、DALL·E 2）的熟悉程度后，我们围绕生成式人工智能是否以及如何适用于不同职业阶段（培养未来科学家、

开展研究、传播研究发现）以及科学方法的各个环节（选择研究问题、形成假设、设计实验、收集数据、分析数据）提出了一系列问题。最后，我们询问了一组问题，以探究在他们的领域使用生成式人工智能可能引发的担忧或弊端。

访谈提纲如下：

背景问题

- 同意参与访谈并录制视频通话
- 人口统计信息
- 研究领域
- 与 Alphabet 的所属机构关系及其他所属机构

了解学科

- 请用不超过三句话举例说明你正在应对的一项科学挑战，让非专业人士也能理解。
- 你认为你所在领域面临的最大挑战是什么？
- 你所在领域的科学家/教员花时间做哪些事情，而你希望这些事情能够自动化？
- 你所在领域的博士生/博士后/青年学者面临什么挑战？

围绕生成式人工智能进行头脑风暴

- 您熟悉生成式人工智能模型吗（例如 GPT-3 或 ChatGPT 等语言模型，DALL·E 2 等图像模型）？ 如果熟悉，请说明您使用这些模型的经验程度。您曾使用过这类模型吗？（如果没有，访谈者将解释什么是生成式人工智能。）
- 您能否设想生成式人工智能工具（无论是现有工具，还是未来更强大的工具）可能会对以下方面产生哪些影响：– 您所在学科的本科生或研究生教育？ – 了解您所在学科或其他学科的相关工作？ – 您确定研究问题、提出或检验假设的方式？ – 实验设计、数据收集或分析实践？ – 研究发现的交流或传播？

对人工智能的担忧

- 要让您信任科学领域使用的生成式人工智能工具，需要具备哪些条件？
- 您对生成式人工智能在您所在学科或整个科学领域中的潜在用途有哪些担忧？
- 您还有其他想分享的意见吗？

3.3 定性数据分析

访谈笔记（以及必要时由人工智能软件自动转录的原始录音）采用定性数据分析技术进行分析。所有分析均由一位研究人员完成（即进行访谈的同一位研究人员），该研究人员在定性方法方面拥有二十多年的专业经验。

研究者采用归纳分析 [50]，通过开放式编码 [12], 反复审阅访谈笔记和记录稿，以归纳出一组主题。随后，研究者利用这些主题进行亲和图分析 [25], ，将响应映射到各主题（并据此合并或完善部分主题类别）。研究发现部分围绕这些归纳出的主题展开，并引用参与者的响应来说明各主题。引号表示逐字引用，其他响应则为概述。

4 研究发现

多个主题贯穿参与者的响应。受访者提到了生成式人工智能在科学工作多个方面带来的各种机遇和顾虑。在接下来的章节中，我们将讨论访谈中出现的几个主要主题，包括教育、数据、文献综述、编程、发现和交流。我们还将讨论使用生成式人工智能可能如何影响人们对科学研究质量的信任。

4.1 教育

受访者一致认为，生成式人工智能的发展可能会改变科学领域本科教育和研究生教育的诸多方面。其中许多变化可能带来积极影响，例如教学方法的创新，以及降低英语学习者的学习门槛。另一方面，许多参与者也对人工智能辅助作弊对评估方法和批判性思维的影响表示担忧。

4.1.1 教育：生成式人工智能带来的机遇

参与者认为，生成式人工智能有许多助力科学教育的机会。生成式模型可用于制定新颖且吸引人的教案。例如，教授可以让语言模型提供一些构想：在讲座中用来阐释定理的现实案例，突出某个概念的作业题，或其他启发性的点子。教师可以从中挑选内容，用于编写或更新讲座内容、实验活动、作业或考试。人工智能模型还可能有助于开发新型演示或为学生打造互动体验。例如，P2 描述了他如何在一门本科地球科学课程中使用 OpenAI 的生成式图像模型 DALL·E 2，让模型生成想象性插图，展示自然灾害发生后世界某些地区可能呈现的样貌（例如“地震后的波士顿”）；这类素材可以丰富课堂讨论与辩论。

另一些人看到了语言模型创造一类新型智能辅导系统的机遇。这类工具可以为学生的作业提供反馈，详细解释学生最初的答案为何错误，并帮助他们自行得出正确答案（而不是通过自动重写作业来“作弊”）。P11 提议，学生可以利用语言模型进行反思式问答，以检验自己对科学概念的理解：学生可以向聊天机器人提问，了解某个概念，然后再向机器人解释该概念，以检验自己的理解深度并构建心智模型。P16 同样认为，能够与学生进行教学对话、帮助他们自我测试对某个主题的知识掌握情况的人工智能机器人具有价值。

语言模型还可以帮助弥补教员和助教（TA）有限的时间资源。学生遇到问题时，可以先向聊天机器人提问；只有在聊天机器人无法提供充分帮助时，才需要去助教或教员的答疑时间。P10 指出，学生不仅可以在求助助教之前向聊天机器人提出与课程内容相关的问题，聊天机器人还可以回答与操作流程相关的问题，例如如何排查解决数学、统计或仿真软件的问题，而这些软件通常是完成大学科学课程问题集所必需的。P4 指出，能够向科学聊天机器人询问高级概念（例如“为什么拓扑绝缘体是热门研究领域？”），可以让学生获得相关信息；否则，如果他们无法轻易接触到某一专业领域的专家，他们的问题可能就得不到解答。

科学术语可能令学生望而却步，成为学习的障碍。P16 指出，语言模型可以通过让学生询问特定术语或概念的含义来帮助他们理解，从而改善对课程材料的理解，减少这种阻碍。P7 也提到，她所在的领域（生物化学）存在严重的“术语障碍”，语言模型不仅可以帮助学生理解阅读材料中的行话，也能在课堂上提供支持。例如，研讨会演讲者提到某种蛋白质时，学生可以迅速查询该蛋白质的性质和相互作用，以便理解讲座接下来的内容。

最后，受访者指出，科学领域的许多学生写作能力并不强，而能够协助写作的语言模型将提高大量学生的写作质量。

P11说：“我知道人们对 [ChatGPT] 有很多负面反应，教授们也非常担心 [新形式的] 抄袭，而学生们因为 [不写作]而缩短了思考过程……但实际上，我认为 [语言模型] 将在很多方面提高底线，帮助那些写作有困难的人……优秀作家仍然有发挥空间， [模型] 不会替你完成所有工作。”P19指出，他所在系（工程学和地球科学）的研究生中，超过一半是在国外出生、以英语为第二语言的学生；他认为，语言模型“将有助于减少在发表和写作方面存在的不平等，让[英语学习者]能够更好地发表和写作”。

4.1.2 教育：对生成式人工智能的担忧

不出所料，作弊是参与者在思考生成式人工智能将如何影响教育时经常提及的话题。参与者普遍认为，人工智能辅助写作工具降低了作弊的门槛（例如生成书面论文或回答作业题），也使作弊更难被发现（即由于生成内容的流畅性）；但对于应担忧到何种程度，参与者之间存在分歧。

对 AI 辅助作弊不太担忧的教员大致可分为三类：

第一类人认为（借用歌手泰勒·斯威夫特的话）“作弊者终究会作弊”——也就是说，想作弊的学生可能本来就已经在作弊，而没有作弊倾向的学生不太可能因为大语言模型的出现而受到诱惑。一些教员认为，生成式人工智能工具会像其他技术一样，起初引发人们对教育受损的恐慌，但最终会被接受为有价值的工具；P17 认为，人工智能最终会被视为“像计算器一样”，不过也指出（正如使用计算器时一样），教育工作者和学生需要认真思考，在什么情况下以及应如何在有无该工具提供的辅助框架下完成任务。同样，P9 强调，教学生如何有效使用生成式人工智能将十分重要，并将其类比为学习使用搜索引擎：“也许这是一种你需要学习使用的工具，20 年前我还不知道如何使用 Google 搜索……现在我已经掌握了搜索自己想找的东西的技能……”事实上，P13 指出，如果教员不允许学生将人工智能用作教育工具，那将是“虚伪的”，因为科学家自己也看到了利用人工智能作为工具来增强自身工作流程的价值。

第二阵营将 AI 辅助作弊的潜在可能性视为一个机会——某种程度上，这是对教员的警醒，促使他们将评估方法改造为可能更有效的形式。P19 认为，提高口试的使用比例（这已经是许多科学领域研究生的常见做法）将是改变评估实践的一种方式。

第三阵营指出，AI 辅助作弊可能会成为本科低年级课程所关注的问题，但对于高年级本科课程或研究生而言，这种情况发生的可能性要小得多，因为当前的大语言模型缺乏领域专门知识，无法为更高阶的科学课程生成准确答案。P10 提出，也许有一天，未来的人工智能工具能够达到“三年级研究生”的知识水平，但他认为，这些工具永远不会具备资深博士候选人所拥有的高度专业化领域知识（资深博士候选人的工作是创造新知识）。除了缺乏在高阶科学课程中成功作弊所需的足够领域专门知识之外，教员还指出，即使是在精英大学的本科低年级课程中，人工智能工具的写作流畅度可能也不足以支持作弊，因为这些大学学生的作业标准相当高；例如，P20 指出，ChatGPT 生成的文本水平相当于精英大学中“成绩为 C 的学生”。

尽管一些受访者并不特别担心使用人工智能工具作弊，但许多受访者认同这样一种观点：AI 辅助作弊是个问题，而且会让作弊达到前所未有的规模（降低在各种作业类型中作弊的门槛，例如课后习题集、论文、编程等），并使此类作弊更难被发现。P8 指出，虽然在理想情况下，教员会花时间设计无法通过人工智能规避的新型作业和考试，但在现实中，教员已经不堪重负，许多人“抽不出 [时间]来编写……新的考题”。AI 辅助作弊可能会在科学教育中造成不公平的竞争环境，也让教员难以评估学生对概念的真实理解。学生广泛使用生成式人工智能（或人们认为学生广泛使用生成式人工智能），可能会降低学生学习动机（因为他们会认为，作弊率高使考核本身有失公平），也可能降低科学奖项和发表平台的声望（例如，如果有人怀疑某个奖学金申请是由人工智能撰写的，那么从某个特定机构获得研究生或博士后奖学金，可能会被视为荣誉较低）

如果有人怀疑申请会士资格的提案可能是由人工智能撰写的，那么当选为科学团体成员或许会被视为一种较低的荣誉。
另一个担忧是，即使使用生成式人工智能工具最终成为常态，在技术层面上也不再算作课程作业中的“作弊”，学生仍会因此错失自身教育中的关键部分。例如，P8指出，使用人工智能辅助工具总结课程阅读材料中的观点或撰写论文初稿，可能会剥夺学生进行批判性反思的机会；而能够轻易使用语言模型的学生，也会失去花时间进行概念综合这一“创作过程”的动力。

除了对作弊的担忧外，一些参与者还担心，即使学生以正式获准的方式使用生成式人工智能，例如接受辅导或出于好奇进行问答，也可能接触到事实性错误。对于年级较低的学生（即本科生），这尤其令人担忧，因为领域专门知识较少的人可能较难识别语言模型生成的幻觉、虚假信息或其他误差；P12指出：“在教育方面……人们完全可能被引向错误的方向…… [如果] 他们向其学习的人知识不够扎实。”除了明显的错误信息外，另一种教育风险是，学生可能不加质疑地接受语言模型提供的信息，而不考虑模型中的偏见可能使他们只接触到关于某个科学问题的多种观点中的一种。P14将这种风险称为“摘要生成的政治性”，并指出，模型在总结某个科学主题时选择强调或省略的内容可能影响讨论。P2指出，科学家自身往往也会对事实存在分歧，科学认知也经常随时间演变；他强调，人工智能究竟是在教导有志成为科学家的学生了解“主流观点”，还是教他们“如何超越主流观点”，两者之间存在微妙差别。

4.2 数据

数据是科学方法诸多方面的核心。数据可以代表观测；对这些观测进行分析有助于提出假设。为检验假设而设计的实验又能生成数据，对这些数据进行分析有助于得出结论，并不断调整预测、实验等。参与者认为，生成式人工智能技术在数据采集、准备和分析方面有许多应用机会，但他们也担心数据集中可能引入误差、数据分析中人类反思的缺失，以及现代机器学习方法是否适用于“小数据”科学领域。

4.2.1 数据：生成式人工智能的机遇

在访谈中，最常出现的主题之一是生成式人工智能有望提高数据准备和使用的效率，和/或创建新的数据源。科学家在众多领域中最常提到的一项挑战，是需要投入大量劳动来汇集数据集并为使用做好准备。这包括合并由不同来源整理的相关数据（例如，来自多个政府机构的天气卫星观测数据）；这需要先找到每个数据源并获取访问权限，然后将数据转换为兼容的格式（因为每个来源通常都有自己的元数据标准）。除了合并数据外，数据清理也十分繁琐且耗时，需要删除错误、不完整或重复的数据。最后，在许多学科中，数据标注任务目前需要耗费时间和资金，标签可能由专业人工标注员生成（例如，言语病理学家标注声学信号的特征），也可能由付费众包工作者完成较不专业的任务（例如，转录言语内容）。研究生可能要花数月甚至数年汇集和准备数据集，之后才能着手对数据进行科学分析。

例如，P19（一位地球科学教授）指出，他所在领域的数据通常是公开的，但“数据公开并不意味着数据就能使用”。他将研究生在该领域必须完成的大量数据整理和数据清理工作称为“研究的核心瓶颈 []”。如果能够用自然语言向人工智能模型发出查询，让它完成数据集准备中的部分或全部步骤（或者至少生成能够处理数据处理步骤的代码），将极大提升科研效率。

许多领域的数据量和异质性，使许多科学家难以识别有用模式。P20（其专业领域是环境遥感）指出，

由于卫星图像的可用性，她所在领域的数据量如今前所未有。P9（一位气候科学家）表示，每天都会根据传感器观测、模拟和实验生成数拍字节的气候数据。P18（一位生物信息学家）观察到，在精准医学领域，需要测量的时间点和指标非常多，因此必须进行“数据缩减”，以限制科学家在分析中能够考虑的信息。能够大规模识别大型异构数据集中的模式的人工智能系统，或许能在许多领域带来新的洞见。这类系统可以提供界面，让科学家使用自然语言查询数据；例如，P12（一位开发量子计算硬件的研究人员）指出，量子硬件专家目前依靠“直觉”来调试故障硬件。他们能够访问数据，但由于数据异质且分散，数据格式并不便于查询和决策。

两位社会科学家（P13，社会学家；P17，医学人类学家）指出，生成式人工智能系统有望加快并规范定性数据分析的流程。定性数据分析（例如主题分析、开放式编码、扎根理论等）通常耗时且劳动密集，需要经过大量培训才能精通某种特定方法，而且很难准确描述和复现。P13认为，语言模型或多模态模型有望“成为我们 [社会科学] 数据分析工具箱中的一种方法”，但也指出，要让学界接受这一做法，就必须提高透明度和可解释性，说明人工智能做了什么以及如何做到：“我们以自己的方法论为傲……如果我们无法充分说明自己的 [数据分析]方法……就无法据此发表研究。”P17强调，采用更结构化、自动化的定性数据分析方法，可以提高科研机构利用定性数据的程度（他提到近期与美国食品药品监督管理局的一次交流，该机构在处理这类数据时遇到了困难）；这有助于确保定性分析不会相较于定量分析而“被降格为一种工具”。

生成式人工智能模型也能通过组建全新的数据集，为科学探究提供价值。这可以通过生成合成数据来实现；例如，P20指出，可以创建合成图像数据，通过对卫星因云层覆盖或数据收集受到政治限制等原因而无法直接捕捉的数据进行建模，从而推动地球科学发展（例如，无法让科学卫星飞越俄罗斯或中国）。除了合成数据，人工智能还可以利用语言模型生成交互式提示词，支持从人类受试者收集数据，从而助力数据集创建。例如，P11指出，语言模型可用于创建刺激材料，以便从终端用户收集结构化语音数据；如果人工智能代理采用对话方式引导参与者说话（而不是要求参与者简单重复固定的语音提示），这类数据可能比许多现有的数据收集方法更真实。另一种利用人工智能开发新数据集的方法，是让基于科学语料库训练的语言模型爬取一批论文，提取目前可能分散在大量来源、格式和详略程度各异的资料中的数据，并将其转化为可用的数据集。例如，P7谈到，在实验室合成蛋白质“有点像一门黑箱技艺”（博士后研究人员必须优化多个参数，例如用哪种细胞来制造蛋白质、将细胞置于哪种培养液中、温度设定多少、持续多长时间等）；随后，从可靠的研究论文中提取常用方法的数据集，就能让研究人员用自然语言向模型提问，并获得这类任务的推荐参数。

某些领域的数据量庞大，需要压缩数据，才能将其存储在大学科学家能够负担得起的基础设施中。现代机器学习技术或许可以解决这一问题：如果基于相关数据构建模型，那么只存储嵌入空间表示，而不存储原始数据点，可能是一种有效的压缩方法。P6认为，嵌入表示或许能解决这一“数据传输”问题，让科学家不必为存储支付“50万美元”；不过他也指出，“传统派”可能会反对这种表示方式，因为在他的水文学领域中，他们无法为每个变量标注其物理量。P2也表达了类似的想法，指出复杂物理现象的机器学习模型或许是最精简的表示方式。他认为，这类模型本质上难以理解，因为它们用数万亿个参数来表示复杂系统；但如果黑箱计算科学模型的预测足够准确，即使科学家无法完全理解模型，P2也认为使用这些模型是可以接受的。

4.2.2 数据：对生成式人工智能的担忧

尽管参与者普遍看好生成式人工智能改善科学数据交互诸多方面的潜力，但他们也对这种热情提出了一些保留意见。例如，P4很期待语言模型能够从学术文章中提取非结构化信息，以创建新的数据集；但她指出，这些数据必须“值得信任”，并担心使用语言模型处理研究论文以提取数据，可能会给最终数据引入误差。P11认为，生成式人工智能有望加快缓慢且成本高昂的数据标注流程，但他也对人工智能标注复杂数据（例如具有“1000种不同声学特征”的语音数据）的准确性表示担忧（不过他也承认，人类专家生成的标签“准确率也达不到100%”）。

尽管 P17 认为语言模型有望为定性数据分析技术带来标准化和效率，但她指出，完全自动化定性数据分析会抹去人类反思的某些方面，而这些方面被视为人类学实践的根本。

P11（一位为沟通障碍人士开发工具的言语语言病理学家）指出，“小数据”是其专业领域面临的一项挑战（例如，患有某些沟通障碍的人相对较少，如肌萎缩侧索硬化症），并疑惑生成式人工智能模型（通常基于大型数据集训练，数据参数往往达到数十亿甚至数万亿）是否适用于他的专业领域。类似地，P2（一位从事地震预测研究的地球物理学家）指出，由于大地震是罕见事件，可能没有足够的数据来训练机器学习模型对其进行预测（而且，要验证此类预测是否准确，可能需要数十年甚至数百年）。

最后，P2 对使用生成式人工智能制造虚假数据表示担忧，并指出，在生命科学等领域，虚假数据尤其是一个风险，因为这些领域通常不会开放共享数据（与地球科学、物理学和天文学等以开放数据共享为常态的领域相比）。封闭数据集会使虚假数据更难被发现，因为同行无法检查这些数据。P2 担心有人会生成虚假数据（例如 PCR 数据、基因组数据或蛋白质结构数据），以吸引生物科学初创公司的资金，和／或支持虚假的“发现”，从而美化自己的学术声誉。

4.3 文献综述

了解本领域的科学文献动态是科学家工作的重要组成部分。这之所以重要，原因有很多，包括通过持续关注重要趋势和发现来保持专业技能与时俱进、从中汲取灵感以构思未来的研究方向、寻找潜在的合作者，以及在开展新的研究项目之前进行文献综述，以验证研究成果的新颖性、在先前知识的基础上恰当推进研究（例如，采用与过去实验相当的方法和数据，以确保可重复性和可比性），并在学术出版流程中恰当地引用相关工作。生成式人工智能，尤其是语言模型，即将大幅改变文献综述流程的诸多方面。

4.3.1 文献综述：生成式人工智能的机遇

随着科学发展的速度和规模不断提升，科学家越来越难以及时掌握最新进展（并从低质量文章乃至科学垃圾信息中筛选出值得阅读的高质量文章）。参与者认为，人工智能工具若能根据感兴趣主题的自然语言描述，推荐相关文章，将非常有价值（例如，关于某个特定子领域、与某个具体假设相关，或采用某种特定方法论的文章等）。P2指出，这类工具有助于研究生快速熟悉一个研究领域；他提到：“这一代学生不会把周末都用来读旧文献……”并认为，帮助学生发现相关文献的新方法将使他们受益。P3也认同语言模型能够帮助科学家快速了解新主题，并指出科学家开展新项目时必须消化大量文献：“你只能一篇接一篇地读论文……”同样，P6指出：“科学文献的数量太庞大，已经难以发挥作用。”P18谈到，现有的学术搜索引擎难以应对她所在领域（干细胞生物学和生物信息学）的海量文献，并举例说

搜索一种特定蛋白质 P53 时，结果可能多达数千万篇研究论文（因为 P53 与一半的癌症有关）。

根据查询生成一份可供阅读的相关工作清单固然有用，但P5等参与者指出，语言模型应将建议的引文直接链接至原始文章，以便科学家核实其真实性，并直接查阅原文。此外，参与者表示，他们希望获得的不仅仅是一份参考文献清单，还重视有关每篇建议参考文献为何被认为与其查询相关的信息。例如，P7表示希望了解一组引文之间的关联；P13也希望人工智能能够解释为何建议某条特定引文，以便他决定是否值得花时间阅读全文。

除了检索（以及概括和综合）学术文献外，P12还指出，生成式人工智能工具有望帮助科学家发现一系列相关在线资源，并可能根据其专业知识或教育水平进行针对性推荐。例如，针对某个主题，本科生的查询可能得到YouTube视频或幻灯片，而研究生的查询结果除了原始文章外，还可能包含研究研讨会的录像。多模态内容（幻灯片中的图表、视频片段）也可以穿插在文本中，为不同能力水平的科学家生成引人入胜的领域交互式摘要。从事量子算法研究的P16指出，由于量子计算是一个非常年轻的领域，即使是入门主题的知识，也可能需要阅读一篇“40页的研究论文”才能掌握；他强调，人工智能工具有望以更易理解的方式为非专业人士或新入学的研究生总结概念，从而减轻其学科中“解读文献所带来的异常沉重的负担”。

P4指出，除了找出一组值得阅读的文章外，能够根据对文献的解析，推荐在特定主题领域开展工作的科学家的模型，也能为研究生或其他刚接触某个新领域的人提供帮助。这有助于科学家了解某个子领域的关键人物和机构，找出自己希望持续关注其研究的人，甚至推荐合作伙伴。P13指出，他所在领域的青年学者需要了解“哪些社会学家群体彼此合作、共同发表文章、共同思考，而你会希望成为其中一员”。他还指出，文献综述工具可以帮助青年学者找到适合发表自己研究成果的平台，例如了解哪些人在特定期刊上发表文章、哪些人引用这些期刊上的研究等。

P6提到，在科学领域，研究者常常形成彼此引用对方研究的“孤岛”（有时受地理因素影响）；如果人工智能能够找出那些在自身常见引文图谱之外开展相关且高质量研究的科学家，就能打通这些孤岛，并有望推动更多科学进展。P13指出，除了让科学家了解常见引文网络之外的研究，这类工具还可能积极推动多元、公平与包容议程，例如提高来自历来处于边缘地位的人口群体（如女性、黑人、拉丁裔等）以及来自更多不同机构（如历史上的黑人大学）和／或全球南方的学者的知名度。此外，P13还指出，由于付费墙的存在，资源较少机构的科学家无法获取许多期刊论文；能够总结此类文献的人工智能工具可以扩大科学知识的获取范围（不过，通过人工智能摘要绕过付费墙也会带来一系列自身的伦理与知识产权问题）。

几位参与者指出，大语言模型能够通过为文献综述增添价值，而不仅仅是推荐文章，还可以实际生成某一主题的摘要，从而加快对新领域的学习；正如 P1 所描述的，这超越了搜索引擎的功能，真正实现了“以可用形式汇集知识”。他设想有一种智能助手能够为科学家“阅读文章并进行摘要”，这样科学家便可以进一步判断哪些文章值得投入时间完整阅读。同样，P7 也希望有一种工具，能够简明扼要地总结任何特定学术文章的要点。

P3 和 P4 都表示，希望有工具能够帮助综合一组文章并从中提炼洞见；P3 指出，这些联系“可能仅仅因为科学家 [进行大规模文献综述]时的带宽限制而被忽略”。P1 的设想更侧重于对单篇文章进行简短摘要，而 P3 的设想则更侧重于压缩一批科学文献（他形容医学文献“规模庞大”）；他设想自己可以在 PubMed 上搜索某个主题，获取排名前 200 的搜索结果，然后让人工智能“给我一份两页的摘要汇编，概述这些 [200]文章的内容”。

P3 指出，人工智能不仅有机会总结文章，还可以按需生成定制综述文章。综述文章刊载于许多领域的学术期刊中，通过考察大量研究成果来概述并综合最新进展。然而，正如 P4 所指出的，综述文章很快就可能过时（在许多科学领域发展迅速、而期刊出版流程相对缓慢的情况下，有时甚至在文章发表时就已经过时）。此外，正如 P3 所说，有些主题可能根本没有综述文章；他在为新项目做准备、熟悉相关文献时提到：“你也会读综述论文…… [但]有时并没有这类综述文章。”P12 也表示，希望获取一组与“小众主题”相关的参考文献，而现有摘要汇编很可能不会涵盖这类主题。P13 指出，社会学家“无法读完 [自己领域]里的所有内容”，因此特别依赖“年度综述文章”：由知名资深研究人员“就 [自己的研究主题]……撰写一份领域综述”，而“我们所有人都会读”。书评在社会学领域也尤为重要；P13 描述说，知名社会学家会阅读彼此的书籍并撰写书评，这些评论因体现了领域内其他人的看法而备受重视。能够撰写此类年度综述文章和书评的人工智能工具，或许能为其他社会学家带来价值；不过，人工智能撰写的综述文章是否能以其及时性和专业性，胜过科学家希望通过业内特定资深人士的独特视角来获得提炼后摘要的需求，仍有待观察。

用于支持文献综述的语言模型不仅可以总结单篇论文或综合一组论文的研究发现，还可以发现一个领域随时间变化的模式，揭示对该领域的新手（甚至是资深科学家）而言可能并不明显的宏观趋势。P15 表示，希望有人工智能工具能够浏览一批新近研究成果，并告诉他某个学科中“什么最热门”。此外，这类大规模分析还可能发现“文献空白”（P7），从而指明未来探索的机会。P18 提出了相关设想：他希望能够向人工智能询问“这一领域目前已有何种认识？”，以节省时间，避免探索文献中已有充分研究的研究问题。

除了根据科学家用自然语言描述的兴趣，为其推荐主题相关的论文外，P4 还指出，如果人工智能系统能够从文献中的过去模式中学习，并据此预测新发表论文可能产生的未来影响（即哪些论文未来可能会被高被引、赢得奖项等），将会很有价值；这有助于应对不断增长的文献带来的跟进难题，将科学家的注意力引向重要性概率较高的内容。P8 指出，大量低质量科学文献的激增加剧了跟进相关研究发现的难度：“有太多期刊根本不在乎发表内容的质量”；她认为，人工智能工具可以帮助识别“优秀研究成果”，引导科学家有限的注意力资源。

正如参与者所描述的，要及时了解自己科学专长所属子领域的最新文献颇具挑战。紧跟自身领域的进展已让大多数科学家无暇顾及其他科学学科中的相关发现和想法。然而，跨学科联系对科学发现可能极具价值。例如，P3指出，这类联系有时能为科学带来一次“量子飞跃”，并提到物理学中的概念如何显著推动计算生物学领域的发展。能够识别不同领域间相关模式或相似之处，并主动推荐跨学科相关文献的人工智能工具，或许能促成意外发现。同样，P15提到，他们大学的系专门寻求聘用在多个领域交叉处工作的教员，因为他们认为这些交叉领域蕴藏着推动科学进步的巨大潜力；他将目前判断一个领域的方法或工具是否适用于另一领域的过程称为某种“随机游走过程”，并指出，如果有一种人工智能能够扫描整个科学界的信息，识别某一领域的进展可能如何应用于另一领域的潜在联系，就会带来巨大的价值。P5（一位材料科学家）认为，跨学科检索相关工作的新方法很有价值，例如用自然语言向人工智能询问一组材料，然后获取与之相关的一组研究领域（例如，了解到某种材料组合“既受到太阳能电池研究人员的关注，也受到生物医学研究人员的关注”，这可能会让人洞察此前未曾考虑过的这些材料的某些特性）。P14指出，目前开展跨学科文献检索颇具挑战，因为需要弄清各个领域使用的关键词，“

同一个概念在不同的 [领域]可能有不同的称呼”；能够解决这一词汇障碍的人工智能工具也可以加速跨学科工作。
几位参与者表示，希望语言模型能帮助他们撰写（或至少起草）科学论文中的文献综述部分。例如，P6设想，可以在人工智能工具中输入一个研究主题，然后获得一篇完整写好的文献综述或相关工作部分。P9希望人工智能工具能自动撰写许多基金申请所需的文献摘要。我们将在关于人工智能辅助的科学传播一节中进一步讨论这一可能性。
语言模型还可以提供交互式阅读辅助工具，供读者阅读某篇文章时使用。例如，P16指出，如果论文提到某个不熟悉的概念（或专业术语），学生可以请模型解释该术语，以帮助理解文章（而无需额外进行文献检索）。

除了通过文献综述了解新主题或跟进自己专业领域的最新进展外，参与者还描述了进行专题文献综述的情况：他们会从大量研究论文中手动提取非结构化数据，例如实验方法或模拟参数的具体细节。例如，P7描述了阅读一系列论文，以找出其他科学家在实验室中合成某种蛋白质时采用的最佳条件，作为设计自己实验的前期准备。她还描述了另一次为实验设计收集信息时进行广泛文献检索的经历；当时她“只是手动创建电子表格，从文献中整理引文，记录 [其他科学家] 做了什么实验、得到了什么结果……涵盖100多篇不同的论文，并提取这些确切的数值 [她感兴趣的某项具体指标]。”这种结构化数据提取过程可能极其耗时，可以通过人工智能工具实现自动化。

4.3.2 文献综述：对生成式人工智能的担忧

在学习科学文献时使用生成式人工智能，幻觉是一个主要顾虑。P2提到，他曾尝试使用 Meta 的 Galactica AI [49]（该模型于2022年短暂发布，并被称为“科学语言模型”）来询问一些地球科学的基础问题，并形容其回答“糟糕透顶”，其中包含错误信息，例如声称“地震是太阳系中最强大的力量”，或错误地说珠穆朗玛峰由火成岩构成。P11指出，ChatGPT等现状中的大语言模型不会提供参考文献链接，这一点“是个巨大的局限……作为学术界的一员，你就是不能把ChatGPT [的话当真”，并指出，手动查找并核实参考文献的需求会抵消大语言模型驱动的文献检索可能带来的时间节省。P13也对幻觉以及人工智能辅助文献综述中的“虚假信息”表示担忧。她指出，由于当前的大语言模型界面与搜索引擎类似（只有一个简单的文本框），最终用户可能会以为自己得到的只是搜索结果，就像使用 Google Scholar 这样的工具一样；这可能会使他们对生成式人工智能输出产生不恰当的信任。在科学家能够相信模型在总结论文时不会捏造科学事实或凭空编造引文之前，事实准确性仍是语言模型成功用于文献检索的一大障碍。例如，有人发现 Galactica 还会编造听起来合情合理、但实际并不存在的论文标题、作者和期刊，有时还会将真实内容与幻觉内容拼凑在一起，例如把一位真实科学家的姓名列为一篇不存在文章的作者。P20讲述了一次经历：她的一位同事曾请 ChatGPT 提供某个主题的一组参考文献，工具返回了一份格式精美的参考文献列表，其中列出了她所在领域的真实科学家姓名和真实期刊，但论文名称却全是虚构的。

P14指出，尽管对单篇或多篇学术论文进行摘要或综合综述能够加快研究进程，但这类摘要在选择哪些内容纳入概要、哪些内容排除在外时，必然会引入偏差。她将基于大语言模型的科学摘要生成称为“一个关键的编辑时刻”，并指出，许多科学家可能会依赖这类摘要，而从不查阅原始资料，因此不仅不会察觉摘要中明显的错误，也不会意识到“摘要生成的政治性”，即人工智能做出的、可能造成解读偏差的决定。另一种偏差风险是，人工智能系统非但没有打破引文孤岛，反而可能通过系统性地将注意力引向特定作品、牺牲对其他文章的关注来强化引文孤岛。

P5 不确定生成式人工智能是否确实有必要用于追踪文献动态；若要在这项任务中采用人工智能工具，取代现行方法（通过谷歌和 Google Scholar 进行检索），就需要明确的指标证明，使用新型人工智能模型能够提升文献检索流程的效率或全面性。这些效率和/或

全面性的提升还必须足以弥补可能失去的关键洞见，而这些洞见来自阅读、反思和综合文献检索结果的过程。P12 对“ [人工智能] 如何塑造我们的认知方式……以及这些摘要会带来哪些损失”表示担忧。不过，由于许多科学家已经在文献检索中使用技术（搜索引擎），相较于科学工作的其他方面，他们在这一用途上尝试新型人工智能系统时可能较少遇到社会阻力。

4.4 编程

OpenAI 的 Codex 和 GitHub 的 Copilot 等生成式人工智能系统 [13], 于 2021年首次推出，专门用于辅助（或可能取代）人类程序员：它们可以生成能够直接运行的软件代码片段，也可以生成供人工开发者编辑的代码，从而加快编程过程。除了专用于编程的人工智能工具外，许多通用大语言模型也能涌现出生成代码的能力 [9], , 因为用于训练模型的许多互联网论坛中都包含代码示例。计算日益成为许多科学领域的一部分；科学工作在各个阶段都需要编程，包括收集或清理数据、运行模拟或其他实验程序，以及对实验结果进行统计分析。大多数科学家没有接受过计算机科学或软件开发方面的正规培训，而对这一群体来说，编写正确的代码（更不用说编写高质量、高效且可复用的代码）极其耗时。能够通过自然语言界面加快或自动完成编程的人工智能工具，具有极大的潜力，可以加速科学发现。

4.4.1 编程：生成式人工智能的机遇 **AI**

几位参与者提到，他们已经尝试使用现有的大语言模型（例如 ChatGPT）以自然语言提出代码编写请求。P16（负责设计量子算法）指出，他曾要求 ChatGPT 编写一个量子算法，“它在某种程度上确实能做到……虽然做得不算特别好，但比我预期的要好。”P11 也提到曾尝试使用 ChatGPT 编程，并观察到“效果不太好……它只是编造出看起来不错、但无法运行的代码”；不过，P11 指出，如果未来的人工智能能够生成准确的代码，那将是“改变游戏规则的”，他估计这会使自己的编程效率“至少提高 30%至 40%”。他还提到，编程占据了其言语语言病理学研究中相当大一部分时间。

P18（一名生物信息学家）描述了编码技能与其领域的紧密交织程度，并指出每个研究实验室都会自行开发定制的科研软件，这是一个耗时的过程：“目前我认为 [生物信息学] 中最大的问题是，每个人都在使用自行生成的系统，每个实验室都建立了自己的服务器、云服务和流程管线……即使只是设置好环境来运行已经建立的软件包，也可能需要数月……”此外，P18 指出，随着学生和博士后离开，许多实验室每隔几年就不得不重新完成这一过程（因为代码往往非常脆弱且文档记录不完善，原作者离开后便无法继续运行）：“当有人离开实验室时，你就失去了关于代码的机构知识[]。”类似地，P10（一名量子化学家）指出，其所在领域一年级研究生面临的一个挑战，是为了使用成熟的量子化学软件包而需要编写多种类型的输入文件；他还指出，现有量子化学代码的文档质量不佳，使这一学习过程更加困难。P9（一名气候科学家）同样指出了由没有软件工程背景的人员编写的代码质量问题，并观察到，气候模型中设计不佳的代码（通常包含数百万行 FORTRAN）导致其他科学家难以使用这些模型。生成式人工智能工具不仅能够编写代码，还能够以结构清晰、文档完善的方式编写代码，从而促进代码的学习与复用；这将消除当今科学领域中大量常见的时间和精力消耗。

更有条理且文档完善的代码不仅有助于研究实验室内部的代码复用，也有助于促进不同机构之间更好地共享代码。P5指出，目前实验室之间缺乏编程标准和代码共享，这正是加快物理科学领域工作进展的一大制约因素。

在计算机科学以外的领域，拥有较强编程技能的科学家寥寥无几；而且，掌握编程技能（无论水平多么业余）本身正日益成为参与科学研究的一道障碍。例如，P2、P6 和 P19 指出，缺乏编码经验是对气候科学感兴趣、希望进入这一领域的学生面临的进入障碍。人工智能辅助编码或许能够拓宽

参与科学领域，尤其是来自历史上代表性不足群体的人士；他们学习编程的可能性较低（例如，自我认同为女性、黑人、拉丁裔等的人士）。
即使生成式人工智能无法生成完全可用的代码，科学家们仍认为，能够通过推荐相关应用程序接口或句法来加快他们自身编程进度的工具很有价值。例如，P11（曾尝试使用 ChatGPT 编程）指出，虽然他对生成的代码质量并不满意，但他发现这类工具有助于了解可用的软件包功能：“比我自己去查快多了”，而且他发现人工智能有助于回答“这个函数怎么用”之类的问题。P11 总结了能够解释应用程序接口的人工智能所带来的价值：“像我这样的人并不是全职程序员，所以我确实经常需要查阅资料。”人工智能在查询应用程序接口信息方面提供支持，也能让经验丰富的程序员受益，因为他们必须学会处理遗留代码系统。例如，几位参与者提到，物理学、气候科学等领域的传统科学系统是用 FORTRAN 编写的，而这门语言通常不会在现代计算机科学课程中教授；能够帮助科学家快速熟悉陌生编程语言的人工智能工具将大有裨益。

P16 还指出，对于做作业题的学生而言，人工智能辅助编码可以节省时间，因为学生目前必须学习复杂的应用程序接口，才能运行模拟或完成其他类型的练习，而编写代码并非这类作业的教学目的。

编写用于数据统计分析的代码（例如，为 R 等软件包编写脚本）是生成式人工智能能够加速科学研究的另一个领域。它不仅可以自动化或加快脚本编写本身，还可以建议适当的统计检验，因为许多科学家并非统计分析方面的专家。P18 指出，人工智能工具的一项益处在于，可以“设置一些防护措施，避免人们使用不恰当的统计方法”。P14 指出，R 和 Stata 等统计软件包使用起来可能“很繁琐”，而通过自然语言请求统计处理代码则能改进工作流程。P3 指出，许多生物学家不擅长进行统计功效计算（而计算不当往往会导致实验设计无法得出明确结论）；因此，支持以自然语言指定统计任务的工具，也有助于实验设计和分析。

4.4.2 编程：对生成式人工智能的担忧

参与者对编程表达的主要担忧在于正确性：如果生成的代码无法正常运行（尤其是程序错误很隐蔽时），科学家就得花时间排查误差，因此未必能节省精力；如果科学家过度信任含有程序错误的人工智能生成代码，还可能导致科学错误。知识产权也是一项担忧：如果人工智能工具提出的代码属于第三方专有代码（因“记忆”错误所致），科学家担心这会让代码的原作者对相关发现提出所有权主张。例如，P19 担心使用人工智能生成的代码“有风险”，因为如果模型曾用他人的源代码训练，其他人可能会对后续工作主张功劳。

4.5 发现

生成式人工智能有望通过改进实验设计、提出新颖的方法和模型，以及支持对实验结果的解读，直接加速新科学知识的发现。本节将介绍科学家对新型人工智能模型如何支持科学发现各个方面的思考，以及需要警惕的问题。

4.5.1 发现：生成式人工智能的机遇

几位参与者提到，生成式人工智能模型有望帮助科学家更合理地分配注意力和资源，例如帮助他们从多个假设中判断哪些最有可能取得成果、需要收集哪些数据来支持探究，或如何优化实验方案和条件。例如，P2指出，监测设备的布设对地震研究至关重要；人工智能模型可以帮助地震科学家规划传感器网络的最佳布局，以最大限度地发挥所收集数据的价值。P6指出，人工智能模型可以帮助为野外水文学家确定最适合进行测量的地点（如今，测量工作往往基于哪些区域容易到达，而不是哪些区域从理论上可能提供最多信息）。同样，P8表示，她对人工智能驱动的工具感兴趣，希望借助这些工具预测某个假设成功的概率，从而确定如何最好地分配资源。

成功的概率。量子化学家P10认为，人工智能工具可以帮助他在某种问题解决方法行不通时，建议接下来尝试哪些方法，并帮助量化多个计算或理论方法中哪一种最合适的概率。

P6 更进一步，猜想比现状先进好几代的人工智能模型，或许不仅能帮助科学家完善假设，还能真正开辟新的探索方向。他评论道：“提出一个问题供人探究，是科学中最难的部分……我完全可以想象，五十年后，人工智能会改变这一流程。”然而，这种看法并不常见——大多数参与者表示，他们认为技术永远无法取代人类科学家凭借创造性洞见提出新研究想法的作用。例如，P12 认为，生成式人工智能工具很难提出新颖的科学研究问题，因为“要向某种机器学习系统提供如此多的背景信息，会极具挑战性……这些信息存在于人们的头脑中，并没有任何可以提供给计算机的具体形式。”P19 指出，从定义上说，研究应该是“从已有事物中进行推断”，这说明在他看来，仅用过去数据训练的模型不可能提出新颖的研究想法。P3 总结了这种关于人工智能在发现新研究问题方面可能发挥的作用有限的看法，并指出，科学家通常必须先了解一个领域的“前沿”，才能决定“哪些问题值得回答”。

P3（一位免疫学家）指出了实验设计的复杂性，尤其是如何设计能够得出明确结果的实验（而不是可能有几种其他假设都能合理解释研究发现的实验）；他还指出，最优设计并非总是可行（例如，实验成本可能过高，或所需时间可能过长）。能够就实验设计提供建议或反馈（包括考虑实际限制），或者帮助科学家从多个潜在设计中选出最有可能带来洞见的方案的人工智能系统，将会是一种很有价值的工具。P3指出，在生命科学领域，Emerald Cloud Lab 等初创企业一直在开发平台，以帮助自动化某些类型的实验设计，但使用此类平台需要创建规范化的模式——若能通过自然语言与大语言模型交互，这类方法将比依赖形式化规范语言直观得多，也更易于广泛采用。

P4指出，在材料科学领域，采用“实验主义”方法进行材料发现通常需要大量反复试错；能够考量广泛实验设计，并基于对该领域既有文献的分析，建议最值得在实验室实际尝试的方案的人工智能工具，可以大大加快材料发现的进程。P16也希望有工具能够“预测实验设计的有效性”，让他可以提出诸如“缓解这项实验测量误差的最有效方法是什么”这样的问题；模型甚至或许还能编写用于运行实验的代码（即在仿真环境中运行）。P18指出，在生物科学领域开展实体（实验室）实验需要“重大投入”；如果能够确信某种实验设计比另一种更有可能发挥效用，就能帮助科学家确定“优先目标”，从而指导他们分配精力。

P7指出，人工智能工具有机会支持生物化学实验设计，帮助研究生规划细节，例如制作实验所需的各种原料用量（如配制一定量、具有特定浓度的缓冲液）。目前，许多学生需要手动创建复杂的电子表格来规划这些细节；能够接收自然语言输入并掌握从研究文献中整理出的方法的计算工具，可以大大简化这一过程。

除了实验室实验之外，P5还谈到，生成式人工智能有望支持基于模拟的研究，例如帮助科学家选择最适合用于探究特定研究问题的模拟设置。P5还提出，人工智能可以通过预测某些性质（如材料科学中的材料性质）来提高模拟效率，从而只需进行部分模拟。

P11 描述了生成式人工智能如何通过支持新的研究方法，在他的领域（言语语言病理学）发挥作用。例如，他可以让语言模型生成一组符合特定标准的言语刺激材料（例如，十个句子，每句至少十五个词）。能够与患者交谈的人工智能聊天机器人不仅可以成为一种新颖的治疗工具，还可以用于收集患者能力变化的数据，并向患者提供有关其言语输出的反馈。

P1（一位行为经济学家）指出，语言模型有可能在某些类型的社会科学研究中取代人类参与者；P1 举例说，他最近读到一篇论文，其中一位经济学家使用现有的大语言模型，在经典的“最后通牒博弈”实验中取代了人类参与者 [2]。这种方法有助于开展新的研究（尽管模型模拟的参与者与人类受试者之间的相关性尚待验证，但这种相关性是 plausible 的，正如过去十年间，在线众包平台上的参与者越来越多地取代实验室参与者，参与许多社会科学研究）；或者，也可以将模拟参与者研究作为一种低成本方法，用于检验社会科学研究发现的可重复性，从而增强心理学等经历过可重复性危机的领域的信心。

生成式人工智能模型有望为研究社会尺度现象的社会科学家提供新的洞见来源，也为研究开辟新的途径。例如，谷歌搜索曾创造了利用临近预报来研判公共卫生趋势的机会，例如流感传播情况（如 Google 流感趋势 [22]），或宏观经济趋势（例如，将搜索失业服务机构地址的人数作为失业人数的领先指标）。本研究中的社会科学家都对如何利用语言模型或多模态模型，洞察经济学、人类学和社会学等领域的社会趋势感兴趣。目前尚不清楚模型的终端用户在多大程度上能够利用模型获取洞见，还是说利用模型获得社会科学洞见需要访问开发者指标的“后端”数据（例如特定提示词或响应的频率）；此外，多数生成式人工智能工具的随机性也可能给将其作为研究和实验材料带来可重复性挑战。

如数据部分所述，能够分析大型异质数据集的人工智能工具，可用于建模复杂现象，并发现规模大到人类难以实际分析的模式。例如，P2 提到，如果能够汇集有关各种现象的大量异质观测数据（如卫星数据、GPS 数据、天气数据等），就有可能预测地震，而无需使用基于物理的模型。同样，P3 指出，如果人工智能工具能帮助科学家从结合人口统计信息、行为数据、病史等的大型数据集中发现模式，精准医学便有望取得进展。

4.5.2 发现：对生成式人工智能的担忧

P5 担心生成式人工智能可能会加剧科学研究中的“研究热潮”。例如，在材料科学领域，他指出，石墨烯首次被发现后，许多研究人员便开始探索其性质，导致相关研究大量涌现，同时也使人们付出了未能探索其他材料的机会成本。如果许多科学家使用机器学习模型来决定下一步探索什么，这可能会加剧此类“研究热潮”；虽然这在微观层面可能无害，但在宏观层面却可能减缓一个领域的进展。P19 表示，他怀疑人工智能会“让大批人群 [科学家] 追逐现成资源，从而催生研究热潮……”（也就是说，选择人工智能工具容易处理的问题或方法）。

准确性和事实准确性仍是将人工智能用于发现过程时的关键顾虑。例如，与大语言模型驱动的聊天机器人对话，未来或许可以成为针对影响言语或认知的各种障碍的一种诊断方法，但 P11 担心“未经人工验证就作出误诊”的风险。P11 指出，必须了解人工智能在不同应用场景中造成假阳性和假阴性的代价，并确保“将人类价值观纳入这一过程”。另一方面，P15 指出，虽然由人工智能错误引发的糟糕科学研究显然并非理想状况，但糟糕的科学研究早已存在——设计或解读不佳的实验、不可重复的研究、发表于低质量期刊的研究等，数量庞大；由新型人工智能推动的科学发现，也需要像由现行方法推动的科学发现一样接受验证和重复检验。然而，现有的大量低质量科学研究可能会降低基于这些文献训练的任何生成式模型的效能，并带来放大无效方法或探索方向的风险；正如 P18 所指出的：“大量科学文献无法重复……这些数据中有很多噪声。”

几位参与者表示担忧，认为生成式人工智能受其本质所限，无法用于发现新知识，因为这类系统只在历史数据上接受训练。如果这类系统只能根据过去的模式进行外推，那么它们是否能预测新概念，事先并不明朗。如果这一观点成立，那么即使在人工智能增强的未来，提出愿景或假设仍将是科学家的独特职责。例如，P5指出

在科学领域广泛使用人工智能工具，可能会降低整体研究的新颖性，使研究更加渐进，因为模型可能更倾向于关注与历史数据高度相关的预测，而不是更具创新性的突破。P6和P9也质疑，基于过去科学数据训练的生成式模型能否很好地泛化并预测未来趋势，尤其是在受气候变化影响的领域。例如，未来的天气模式预计将与过去有相当大的差异。

最后值得注意的是，如果生成式人工智能能够大幅加速科学发现，这不仅可能带来有益于社会的发现（例如新的医疗疗法、清洁能源等），也可能带来潜在的恶意发现（例如生物武器或化学武器）。确保科学人工智能工具具备防护措施，以防止其被用于不道德的用途，是一项重要关切。

4.6 交流

生成式模型在短期内最重要的机遇之一，是支持科学交流，包括撰写学术论文和基金申请，以及制作幻灯片等演示材料。人工智能工具还有机会发挥审稿人的作用，提供初步反馈，帮助学者在正式同行评审之前完善论点。大语言模型辅助科学写作也伴随着诸多风险，尤其是产生出版垃圾信息和科学错误信息的风险。

4.6.1 交流：生成式人工智能的机遇

受访者最常设想的语言模型用途，是让模型生成写作“初稿”，然后由他们进行编辑。学术论文中的“相关工作”等部分尤其被认为适合由人工智能起草，因为许多人认为查找引文既耗时又枯燥（这一观点自然承接了文献综述部分所讨论的内容）。即使是“相关工作”这类相对程式化的部分，参与者也考虑到仍需专家验证写作内容的输出；P3指出，最有用的情况是“它能引用参考文献，这样你就可以去查证，确保这些文献确实支持相关论述”。P8认为，许多科学论文的引言部分初稿也可以交由语言模型撰写，帮助阐明特定研究问题的背景和动机。P9总结了这一共识：大语言模型有助于起草（但不能自动完成）科学论文写作。他表示，自己不相信“按一下按钮”就能生成高质量学术写作，但仍然认可人工智能生成的初稿能帮助他们“提高效率”。

或者，科学家不必让模型先生成初稿，再由自己修改；他们也可以向模型提供非常粗略的草稿（即提纲或一组笔记），由模型将其整理成润色完善、语法正确且条理清晰的正式写作。P16谈到了这种方法的实用性：“现实情况是，很多科学家的写作能力并不强。”

虽然撰写期刊文章和会议论文被视为科学家智力工作的重要组成部分（因此适合部分自动化，但不适合完全自动化），参与者却更希望有能够自动完成基金申请书撰写的工具。许多科学家（尤其是主要任职于大学、而非工业界的科学家）花费大量时间撰写研究基金申请，因此，能够生成完整初稿的基金申请书撰写工具备受期待。P19指出，提交给资助机构的报告中有不少文本“相当笼统”，因此很适合借助人工智能自动化。P20则从略有不同的角度看待人工智能辅助基金申请书撰写工具的潜力：这类工具可以帮助青年科学家学习如何更好地表述自己的想法，从而提高获得研究基金的成功率。

尽管参与者一致认为，他们不会信任语言模型代表自己自主撰写学术文章（即使他们相信模型具备这种能力，也未必希望如此，因为写作过程包含有价值的批判性反思），但他们确实看到了未来几代大语言模型自主撰写风险较低的材料的可能性。例如，P2 期待生成式人工智能能够减轻行政事务负担，比如代他回复大量电子邮件。P19 还指出，他在大学的大部分时间都花在“单调乏味的事情……电子邮件上，其中很多都有通用答案，我相信自动化系统很容易就能生成这些答案”；他还指出，虽然撰写科学论文需要人类专家发挥创造力

，但科学家所做的许多“办公室事务”都可以交由语言模型处理。P1 观察到，面向普通读者的科学发现简明摘要很可能可以完全自动生成（例如，在经济学领域，他指出，媒体报道的股市新闻往往有固定的模式）。

参与者也期待模型能够减少格式处理任务所耗费的时间。例如，P7 提到她曾尝试使用 ChatGPT，将自己为一篇期刊论文撰写的 400 字摘要交给它，并要求将摘要压缩到 200 字以符合字数限制；她认为它完成得相当不错。为了满足字数限制而压缩文本，并为符合不同投稿平台的各项要求反复调整格式，是一项枯燥的工作，科学家们乐于将其自动化（这也是人在回路系统 Soylent 十多年前就预见到的一项任务 [4]）。P14 强调了格式化引文的繁琐（不同投稿平台要求采用不同的引文规范），并感叹许多科学家要花数小时进行枯燥且“看似毫无意义的文风编辑”，而这些工作原本可以自动完成（例如让语言模型输出 LaTeX 标记）。

除了语言模型之外，图像或多模态模型也能在节省科学传播时间方面发挥作用。参与者认为，通过自然语言指令生成科学图表（表格、图表、曲线图、示意图等）的能力很有用，既可用于撰写学术文章，也可用于制作演示文稿和讲座幻灯片。P3 指出：“能够制作图表的工具会很强大。”他感叹制作图表十分耗时，尤其是在数据量很大的情况下，科学家需要想办法用抽象的方式呈现数据，以便传达洞见。P12 希望能够通过多模态交互与模型沟通，比如说“给我画一张像这样的图”，然后提供某种类型的图片作为示例（例如粗略的草图或另一篇文章中的图片）。P14 则设想将自己的数据提供给模型，再让它“创建一张包含人口统计信息的表格……并且我要它采用 [特定的] 格式”。

能够根据一组输入生成包含文本和图表的幻灯片演示文稿，也将极具价值。这可以是面向学术受众的演示文稿（将期刊论文和/或数据作为模型输入），也可以是面向学生的教学讲座（将课程计划或一组多样化材料作为输入，并生成适合受众的幻灯片）。

能够将一份文档（无论是文章、幻灯片等）调整得更适合特定受众，也将是语言模型一项备受重视且节省时间的应用——例如，要求模型针对具有特定背景的受众（科学家、公众、某一学习阶段的学生）或特定场合的限制条件，调整演示文稿或文章（即，制作有特定时长限制的演示文稿，或按特定资助机构或期刊要求的格式生成要点或摘要等）。例如，P4 指出，她可能希望根据目标受众，要求模型重新校准文档中深奥“行话”的程度。P16设想，他可以“给这些模型中的一个一段文字，比如这是一则非常精确的数学陈述，你能不能让它更容易理解一些”；他指出，这在量子计算这样的跨学科领域可能尤其有价值，以便为面向特定受众的展示定制写作，因为不同受众各自偏好不同的术语和技术细节层次（即数学家与计算机科学家与物理学家）。

P8指出，人工智能有机会协助审阅科学文献的流程（即判断是否接收稿件并发表于期刊或其他出版平台）。一种可能是，语言模型（尤其是在该领域既有文献上训练的模型）可以对论文进行初步审阅（即检查其相对于先前研究的新颖性），作为筛选机制，判断论文是否应进入人工审稿人审阅的环节，从而减轻审稿人工作量。语言模型还可以通过提供写作支持来减轻人工审稿人的工作负担，例如根据审稿人对某篇投稿所做的笔记，生成语法正确的审稿意见，或帮助审稿人查找相关参考文献。P14还指出，她所在领域管理审稿人工作量的挑战日益加剧：“疫情之前，出版流程就已经不堪重负……疫情期间情况变得更加困难，很多人无法参与审稿”；她感叹，相较于投稿量，审稿人时间有限，导致研究发现的传播速度放缓。能够减轻审稿工作量的人工智能系统或许有助于解决这一问题（不过，降低撰写期刊论文门槛的人工智能系统也有可能大幅提高投稿率，反而加剧这一问题）。

另外，模型可以在写作过程中，就论文草稿向作者提供类似审稿人的反馈，帮助他们在提交同行评审前改进论文。例如，模型可以就文章向作者提问，甚至尝试自动撰写审稿意见，以帮助作者提前思考如何更清晰、更有说服力地呈现自己的观点。

4.6.2 交流：对生成式人工智能的担忧

P7 尤其担心利用语言模型生成虚假的科学出版物，并蓄意散布有关存在争议（政治化）科学主题的虚假信息，例如疫苗效力、基因编辑或气候变化。她指出，科学文章真实性受到质疑可能导致公众失去信任："科学在社会中确实发挥着重要作用：当事情是对是错时，它能够让人们对此有信心。"

除了蓄意传播虚假信息之外，人们还担心出版垃圾信息（即利用语言模型向科学生态系统大量灌入虚假或低质量论文，以充实作者的简历）；正如 P16 打趣道："人工智能工具可能会让产出相对低质量的科学成果变得更加容易、高效。"P14 指出，出版垃圾信息可能会淹没同行评审流程；P15 则感叹，出版垃圾信息会进一步加剧一项难题：如何决定阅读哪些论文，才能及时了解本领域的进展。P16 也表达了类似担忧：人工智能撰写的低质量文章泛滥，会让人更难"找到你需要阅读的真正有价值的研究成果"。此外，出版垃圾信息还可能带来循环性风险：如果新一代人工智能模型使用人工智能生成的虚假或低置信度信息进行训练，这些信息就可能污染下一代人工智能模型。

4.7 信任

信任人工智能是访谈中的核心议题；参与者经常表达与信任相关的担忧（例如对幻觉和事实准确性的担忧，以及需要引用来源的要求）。访谈结束时，他们还被特别要求思考，要信任生成式人工智能在科学领域的应用，需要具备哪些条件。解决这些问题（有些是技术层面的，有些是社会层面的），对于开发符合科学家和社会价值观的人工智能至关重要，以免危及人们对科学本身的信任。

引文（科学家在自己的写作中也受过如何引用的训练）至关重要。参与者希望能够轻松追溯到原始资料，这不仅适用于人工智能辅助文献检索这类简单任务，也适用于科学人工智能可能支持的各类任务，例如了解人工智能为何推荐某种方法或实验设计、哪些证据支持其假设等。引文不仅必不可少，还应尽可能具体（即指出信息所依据的文章中的具体部分），并且应具有高度可信度和良好质量（即人工智能应优先依据特定可信赖的学者、机构或出版平台所发表的科学成果来提出建议）。

事实准确性和幻觉是参与者特别担忧的问题。描述过与现有生成式人工智能工具（Meta 的 Galactica、OpenAI 的 ChatGPT、谷歌的 Bard）交互经历的参与者表示，这些工具在回答科学问题时缺乏事实准确性，还会捏造不存在的科学参考文献，也会偏离明确给出的指令（例如，P11 曾要求一个大语言模型为言语病理学研究生成一组由十五个单词组成的句子提示词，但系统返回的提示词长度不一，有些只有十一个单词）。P2 还指出，随着科学知识不断进步，事实准确性的定义也会随时间变化；如果模型没有得到频繁更新，就可能无法准确反映当前的科学共识。事实上，由于模型是在历史数据上训练的，可能会偏向较早期的科学观点；P6 打趣说"ChatGPT 是一位物理水文学家"，这是因为他认为，ChatGPT 对水文学问题的回答体现了该领域较为传统的研究，而这些研究已不太能代表当前的最新趋势。

置信度指标被视为帮助科学家判断应对模型输出抱有多大信任的关键。举例来说，P7 回忆起自己曾尝试使用 Meta 的 Galactica 科学模型，但其输出缺少置信度指标令她感到不安（对于她提出的有关常见蛋白质的问题，模型给出的答案完全是幻觉）。人工智能系统若能提供一组排名前 N 的响应（并附带相应的置信度指标），而非仅提供排名第一的响应，也有助于建立对该系统的信任。P12 也提出了相关建议：他希望人工智能能够

明确告知他何时对某事缺乏信心。他指出，当人类科学家对某事没有把握时，可能会给出答案或解释，同时提醒对方最好再查阅一些补充参考资料，或征求另一位专家的第二意见。

可解释性对一些人而言十分重要；科学家希望能够理解任何科学人工智能为何会给出特定的建议或预测。P1 表示：“我可能希望它能够解释自己的答案”，并指出，如果人工智能提出了人类认为不太可能的建议，这类解释就尤其关键。另一方面，P2 表示，在某些情况下，可解释性可能无法实现。他推测，在某些领域，复杂现象的嵌入空间表示或许是最高效的模型，而人类永远无法理解这些表示；如果这类黑箱模型能够准确预测物理现象，科学家就应当使用它们，尽管其运行机制不透明。

交互性（包括交互式可解释性工具）或许有助于建立信任。P6 描述道，借助“交互式仪表板”，用户可以更改模型参数并查看由此产生的结果，这帮助他使水资源管理者相信水文学模型（由更传统的机器学习系统驱动）的准确性和实用性。P12 提出的另一种交互形式，是让人工智能在接收到任务后提供反馈，以确保它正确理解科学家的指令，因为自然语言交互中的细微差异和误解可能会显著改变输出结果。

一些参与者提到，需要制定相关协议，披露人工智能是否参与了科学过程的某个环节，以便其他科学家之后能够以更加审慎的态度看待这类工作。这可能需要开发新工具（例如，判断一篇科学文章是否由人工智能自动生成，以过滤出版垃圾信息），也可能需要制定新的社会规范（例如，人工智能的贡献达到何种程度时，才适合将其列为科学共同作者）。

许多参与者从科学的角度出发，尝试以可量化的方式定义信任。例如，如果科学人工智能系统能够对科学现象作出预测，那么就可以衡量这些预测中有多少是准确的（或许还可以将其与基准进行比较，例如人类专家作出的预测）。这份业绩记录将帮助科学家作出明智判断，决定在多大程度上信任人工智能。P3认为，随着时间推移，凭借准确性建立业绩记录来赢得信任十分重要。他说：“在早期，我会对 [人工智能所说的每一件事]都逐一核查。”P8将此比作科学家自身如何在领域内通过积累出色的工作记录，逐渐赢得信任。P16也将此比作人类必须随着时间推移建立信任，并指出：“我甚至不信任很多同事。这并不是因为我认为他们有恶意，而是因为我共事的人几乎都会在某些时候对某件事产生错误的直觉。”这意味着无论人工智能通过了多少项“信任”测试，科学家也应始终对其保持怀疑态度。

P3认为，人类基线对于比较很重要。例如，对于能够提出实验设计的人工智能，他会想检验人工智能能否提出比自己设计的实验更好的方案。同样，P10设想开展实验，设定某些科学任务，让人工智能和科学专家回答一组问题，然后比较他们的成功率。另一方面，P6认为，与物理模拟（而非人类）进行比较或许就足以令人信服，前提是人工智能模拟能够比这些更传统的模拟系统做出更准确的预测。相比之下，P18提出了更为正式的信任标准——对人工智能技术进行正式临床试验，正如医疗或生物科学领域的其他技术可能接受的检验一样。

P2采取了功利主义视角；他认为，与其关注“信任”本身，不如关注人工智能工具是否具有可证明的效用（描述性效用、预测性效用、节时效用）。

科学界的某些群体比其他群体更容易接受人工智能。在许多领域（包括生物学、材料科学、气候科学等），实验研究者与采用计算方法的人之间已经存在分歧。

参与者认为，在科学领域设计和使用生成式人工智能的过程中，让具备领域知识的人参与其中，对于确保科学准确性至关重要。P8指出，任何科学人工智能都应当“在较长一段时间内让人在回路中……而这些人

应当具备领域知识。”P9指出，尽管他怀疑模型是否可能取代物理学家，但它们或许“能够在某些任务上开展协作……”P11指出，在人工智能的医疗或临床应用中，人工验证对于尽量减少诊断任务中假阳性或假阴性所造成的严重后果至关重要。他强调，需要在科学发现和应用的过程中“融入人类价值观”。

5 讨论

参与者指出，生成式人工智能有许多影响科学领域的机会，包括科学教育、研究和交流。几乎在所有情况下，大家都认为人工智能是一种工具，能够补充、增强或加速科学家的能力，而不是自动化科学实践。

虽然没有参与者认为，像他们这样的领域资深专家很快（或将来任何时候）会被人工智能取代，但 P1（一位经济学家）指出，随着生成式人工智能模型日益强大，许多领域很可能会出现就业替代，并由此造成社会动荡：“人工智能会掏空中层劳动者……”。这或许也适用于科学领域及其他领域，对某些类型的研究生、实验室助理、助教、程序员等的需求可能会减少。无论是人工智能自动化直接导致的社会就业冲击，还是人工智能加速的科学所促成的新科学发现带来的影响，近期和中期内都可能造成严重动荡；即使从长远来看，这些变化最终会加速科学进步、造福社会，并促成就业技能再培训。正如 P1 所说：“我想，回顾百年后的技术……我们不会为一百年前失去的工作感到遗憾。”

现有生成式人工智能工具容易出现严重的事实性错误和幻觉，这目前是这些工具应用于科学领域任何方面的一大障碍。改进模型，并结合广泛的引文及其他可解释性功能，或许能缓解这一顾虑；但准确性对这一群体至关重要，也是未来任何科学人工智能取得成功的关键。

此外，必须认真审视未来人工智能工具可能对科学造成的负面影响；生产力或知识方面的任何收益，都可能被新出现的人工智能引发的弊害所抵消，例如作弊或科学写作过度自动化（进而削弱未来科学家培养过程中的批判性思维能力）、出版垃圾信息（进而影响科学家的注意力并加重审稿负担），以及虚假数据和伪科学（进而损害公众对科学的信任、误导科研经费的分配等）。在设计并考虑是否为这一群体部署任何新的人工智能系统时，必须与领域专家共同审慎评估特定人工智能工具对科学可能带来的利弊。

5.1 局限性

尽管我们的样本包括二十位学科归属各不相同的科学家，但我们的研究并未涉及许多专业领域。此外，所有参与者都在美国生活和工作；国家／文化背景不同的参与者可能会有不同的观点。

由于所有参与者都在某种程度上隶属于 Alphabet，他们很可能（1）比其他科学家（或许计算机科学家除外）更熟悉生成式人工智能，并且（2）可能比同侪更倾向于持技术乐观态度。例如，我们惊讶地发现，尽管参与者都不是计算机科学家，但他们对生成式人工智能都有一定程度的了解，也熟悉已部署的系统（如 ChatGPT）；这可能是他们在 Alphabet 任职期间接触科技行业的结果。

我们没有将计算机科学家纳入受访者范围，因为 Alphabet 的大多数员工对计算机科学领域的知识已经相当普遍；本研究旨在收集更多元的观点。

要推测尚不存在的未来技术能力极其困难。因此，参与者对生成式人工智能潜在应用及相关顾虑的回应，很可能深受现有系统能力以及以下假设的影响：未来

系统只是当今范式的更快、更准确的版本；这种假设很可能并不正确。此外，参与者提出的许多建议似乎并非生成式人工智能模型所特有；某些能力（例如发现数据集中的模式）也可以通过其他类型的机器学习实现，无论使用的是现有系统还是未来系统。

请注意，由于生成式人工智能研究进展迅速，参与者的一些建议和我们的研究发现可能已经过时。这些访谈于2023年1月和2月进行；强大的 GPT-4 语言模型于3月发布后，2023 [41] 已大幅改变现状。

6 结论

我们访谈了来自物理科学、生命科学和社会科学领域的20位科学家，以进一步了解生成式人工智能的进展可能如何影响科学职业。访谈揭示了新型人工智能工具在广泛科学实践中的可能应用，包括教育、数据、文献综述、编程、发现和交流。尽管新兴人工智能技术有望增强科学家的能力并加速知识发现与传播，但仍需保持谨慎并加以反思，以减轻科学人工智能工具可能带来的负面副作用，并防止新技术遭到蓄意滥用。我们正处于一场人工智能革命的前夜，它将影响社会的方方面面，包括科学发现的方法与节奏。

7 致谢

我们谨向访谈参与者致谢，并感谢谷歌 Curie团队提出并支持此次探究。特别感谢 Michael Brenner、Subhashini Venu- gopalan 和 Dan Liebling 提供评论和反馈。

参考文献

[1] Andrea Agostinelli 等。《MusicLM：从文本生成音乐》。2023年。arXiv：2301.11325 [cs.SD].[2] Gati Aher、Rosa I. Arriaga 和 Adam Tauman Kalai。《使用大语言模型模拟多个个体并复现人类受试者研究》。2023年。arXiv：2208.10264 [cs.CL].[3] Emily M. Bender 等。“随机鹦鹉的危险：语言模型会不会太大？”载于：2021年ACM公平性、问责制与透明度大会论文集。FAccT ’21。线上活动，加拿大：计算机协会，2021年，第610–623页。ISBN：9781450383097。DOI：10.1145/3442188.3445922。URL：https://doi.org/10.1145/3442188.3445922。[4] Michael S. Bernstein 等。“Soylent：内置众包功能的文字处理器”。载于：第23届ACM用户界面软件与技术年度研讨会论文集。UIST ’10。美国纽约州纽约市：计算机协会，2010年，第313–322页。ISBN：9781450302715。DOI：10.1145/1866029.1866078。URL：https://doi.org/10.1145/1866029.1866078。[5] Christian Bird 等。“借助 Copilot 起飞：人工智能驱动的结对编程工具的早期洞见与机遇”。载于：Queue 20.6（2023年1月），第35–57页。ISSN：1542-7730。DOI：10.1145/3582083。URL：https://doi.org/10.1145/3582083。[6]Rishi Bommasani 等。“基础模型的机遇与风险”。载于：CoRR abs/2108.07258（2021年）。arXiv：2108.07258。URL：https://arxiv.org/abs/2108.07258。[7] Zalán Borsos 等。《AudioLM：一种音频生成的语言建模方法》。2022年。arXiv：2209.03143 [cs.SD].[8] Tom B. Brown 等。“语言模型是小样本学习者”。载于：CoRR abs/2005.14165（2020年）。arXiv：2005.14165。URL：https://arxiv.org/abs/2005.14165。[9] Sébastien Bubeck 等。《通用人工智能的火花：GPT-4的早期实验》。2023年。arXiv：2303.12712 [cs.CL]。

[10] Carrie J Cai 等。“入职材料：用于开发人工智能助手的跨职能边界对象”。载于：2021年计算机系统中的人为因素CHI会议扩展摘要集。CHI EA ’21。日本横滨：计算机协会，2021年。ISBN：9781450380959。DOI：10.1145/3411763.3443435。URL：https://doi.org/10.1145/3411763.3443435。[11] Minsuk Chang 等。《提示词艺术家》。2023年。arXiv：2303.12253 [cs.HC].[12] Kathy Charmaz。《建构扎根理论：定性分析实用指南（第二版）》。Sage Publications, Ltd.，2014年。[13] Mark Chen 等。《评估基于代码训练的大语言模型》。2021年。arXiv：2107.03374 [cs.LG].[14] Aakanksha Chowdhery 等。《PaLM：借助Pathways 扩展语言建模》。2022年。arXiv：2204.02311 [cs.CL].[15] Alex Davies 等。“通过人工智能引导人类直觉，推动数学发展”。载于：Nature 600.7887（2021年12月），第70–74页。ISSN：1476-4687。DOI：10.1038/s41586-021-04086-x。URL：https://doi.org/10.1038/s41586-021-04086-x。[16]Jonas Degrave 等。“通过深度强化学习控制托卡马克等离子体的磁场”。载于：Nature 602.7897（2022年2月），第414–419页。ISSN：1476-4687。DOI：10.1038/s41586-021-04301-9。URL：https://doi.org/10.1038/s41586-021-04301-9。[17] Victor Dibia 等。《使离线指标与人类对人工智能结对编程工具价值的判断保持一致》。2022年。arXiv：2210.16494 [cs.SE].[18] Nouha Dziri 等。《对话模型中幻觉的起源：问题出在数据集还是模型？》。2022年。arXiv：2204.07931 [cs.CL].[19]Sean Ekins 等。“利用机器学习实现端到端药物发现与开发”。载于：《自然·材料》18.5（2019年5月），第435–441页。ISSN：1476-4660。DOI：10.1038/s41563-019-0338-z。URL：https://doi.org/10.1038/s41563-019-0338-z。[20]Lasse Espeholt 等。“利用深度学习进行未来12小时降水预报”。载于：《自然·通讯》13.1（2022年9月），第5145页。ISSN：2041-1723。DOI：10.1038/s41467-022-32483-x。URL：https://doi.org/10.1038/s41467-022-32483-x。[21]Alhussein Fawzi 等。“利用强化学习发现更快的矩阵乘法算法”。载于：Nature 610.7930（2022年10月），第47–53页。ISSN：1476-4687。DOI：10.1038/s41586-022-05172-4。URL：https://doi.org/10.1038/s41586-022-05172-4。[22] Jeremy Ginsberg 等。“使用搜索引擎查询数据检测流感疫情”。载于：Nature 457.7232（2009年2月），第1012–1014页。ISSN：1476-4687。DOI：10.1038/nature07634。URL：https://doi.org/10.1038/nature07634。[23] Samuel Goldman 等。“利用受领域知识启发的化学式转换器标注代谢物质谱”。载于：bioRxiv（2022年）。DOI：10.1101/2022.12.30.522318。eprint：https://www.biorxiv.org/content/early/2022/12/31/2022.12.30.522318.full.pdf.

网址：https://www.biorxiv.org/content/early/2022/12/31/2022.12.30.522318.[24] Jonathan Ho 等。《Imagen Video：使用扩散模型生成高清晰度视频》。2022年。arXiv：2210.02303 [cs.CV].[25] Karen Holtzblatt 和 Hugh Byer。《情境设计：定义以客户为中心的系统》。Elsivier，1997年。[26] Wei-Ning Hsu 等。《用于可控语音合成的层次生成建模》。载于：国际学习表征会议。2019年。网址：https://arxiv.org/pdf/1810.07217.[27] Ellen Jiang 等。《借助生成式语言模型发现自然语言编程的句法与策略》。载于：2022年 CHI 计算系统中的人为因素会议论文集。CHI ’22。美国路易斯安那州新奥尔良：计算机协会，2022年。ISBN：9781450391573。DOI：10.1145/3491102.3501870。网址：https://doi.org/10.1145/3491102.3501870。

[28] Yucheng Jin、Lanyi Yang 和 Chia-En Chiang。《使用机器学习方法识别系外行星：初步研究》。载于：International Journal on Cybernetics & Informatics 11.2（2022年4月），第31–42页。DOI：10.5121/ijci.2022.110203。网址：https://doi.org/10.5121%2Fijci.2022.110203.[29]John Jumper 等。《借助 AlphaFold 高精度预测蛋白质结构》。载于：Nature 596.7873（2021年8月），第583–589页。ISSN：1476-4687。DOI：10.1038/s41586-021-03819-2。网址：https://doi.org/10.1038/s41586-021-03819-2.[30] Hiroaki Kitano。《诺贝尔图灵挑战：打造科学发现引擎》。载于：npj Systems Biology and Applications 7.1（2021年6月），第29页。ISSN：2056-7189。DOI：10.1038/s41540-021-00189-3。网址：https://doi.org/10.1038/s41540-021-00189-3.[31] Skanda Koppula 等。《运用深度学习刻画大型星系并合的方法》2021年。arXiv：2102.05182 [astro-ph.GA].[32] Chinmay Kulkarni 等。《一字胜千图：作为人工智能设计素材的提示词》。2023年。arXiv：2303.12647 [cs.HC].[33] Aitor Lewkowycz 等。《使用语言模型解决定量推理问题》载于：NeurIPS。2022年。网址：https://arxiv.org/abs/2206.14858.[34]Nestor Maslej 等。《2023年人工智能指数年度报告》。2023年。网址：https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index_Report_2023.pdf.[35] Marcelo C. R. Melo、Jacqueline R. M. A. Maasch 和 Cesar de la Fuente-Nunez。《借助人工智能加速抗生素发现》。载于：Communications Biology 4.1（2021年9月），第1050页。ISSN：2399-3642。DOI：10.1038/s42003-021-02586-0。网址：https://doi.org/10.1038/s42003-021-02586-0.[36] Meredith Ringel Morris 等。《生成模型的设计空间》。载于：NeurIPS 2022以人为中心的人工智能研讨会。2022年。[37] Hussein Mozannar 等。《读懂字里行间：人工智能辅助编程中的用户行为与成本建模》。2022年。arXiv：2210.14306 [cs.SE].[38] 美国国家工程、科学与医学学院。《用于加速发现的自动化研究工作流程：闭合知识发现闭环》。华盛顿特区：美国国家科学院出版社，2022年。ISBN：978-0-309-68652-5。DOI：10.17226/26532。网址：https://nap.nationalacademies.org/catalog/26532/automated-research-workflows-for-accelerate

[39] Tung Nguyen 等。《ClimaX：天气与气候基础模型》。2023年。arXiv：2301.10343 [cs.LG].[40] Mohammad Sadegh Norouzzadeh 等。《用于相机陷阱图像中物种识别与计数的深度主动学习系统》。2019年。arXiv：1910.09716 [cs.LG].[41] OpenAI。《GPT-4 技术报告》。2023年。arXiv：2303.08774 [cs.CL].[42] Pranav Rajpurkar 等。《CheXNet：利用深度学习检测胸部X光片中的肺炎，达到放射科医师水平》。2017年。arXiv：1711.05225 [cs.CV].[43] Aditya Ramesh 等。《利用 CLIP潜变量进行分层文本条件图像生成》。2022年。DOI：10.48550/ARXIV.2204.06125。URL：https://arxiv.org/abs/2204.06125。[44] Rajarshi Roy 等。“PrefixRL：利用深度强化学习优化并行前缀电路”。载于：2021年 ACM/IEEE第58届设计自动化会议（DAC）。2021年，第853–858页。DOI：10.1109/DAC18074.2021.9586094。[45] Chitwan Saharia 等。《具备深度语言理解能力的照片级真实感文本生成图像扩散模型》。2022年。DOI：10.48550/ARXIV.2205.11487。URL：https://arxiv.org/abs/2205.11487。[46] Amir Shanehsazzadeh 等。“利用生成式人工智能解锁从头抗体设计”。载于：bioRxiv（2023年）。DOI：10.1101/2023.01.08.523187。电子预印本：https://www.biorxiv.org/content/early/2023/01/09/2023.01.08.523187.full.pdf。URL：https://www.biorxiv.org/content/early/2023/01/09/2023.01.08.523187。

[47] David F. Steiner 等。“评估联合人工智能与病理学家评估在前列腺活检复核与分级中的应用”。载于：JAMA Network Open 3.11（2020年11月），e2023267–e2023267。ISSN：2574-3805。DOI：10.1001/jamanetworkopen.2020.23267。电子预印本：https://jamanetwork.com/journals/jamanetworkopen/articlepdf/2772831/steiner_2020_oi_2007 URL: https://doi.org/10.1001/jamanetworkopen.2020.23267.[48] Minhyang (Mia) Suh 等. “人工智能作为社交黏合剂：揭示深度生成式人工智能在社交音乐创作中的作用”。载于：《2021年计算机系统中的人为因素CHI会议论文集》。CHI ’21。日本横滨：Association for Computing Machinery，2021年。ISBN：9781450380966。DOI：10.1145/3411764.3445219。URL：https://doi.org/10.1145/3411764.3445219。[49] Ross Taylor 等。《Galactica：面向科学的大语言模型》。2022年。arXiv：2211。09085 [cs.CL]。[50] David R. Thomas。“用于分析定性评估数据的一般归纳方法”。载于：《American Journal of Evaluation》27.2（2006年），第237–246页。DOI：10.1177/1098214005283748。[51] Romal Thoppilan 等。《LaMDA：对话应用语言模型》。2022年。DOI：10.48550/ARXIV.2201.08239。URL：https://arxiv.org/abs/2201.08239。[52] Ashish Vaswani 等。“Attention is All you Need”。载于：《神经信息处理系统进展》。I. Guyon 等编。第30卷。Curran Associates, Inc.，2017年。URL：https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa

[53] Ruben Villegas 等。《Phenaki：基于开放领域文本描述的可变长度视频生成》。2022年。arXiv：2210.02399 [cs.CV].[54]Laura Weidinger 等。“语言模型带来的伦理和社会危害风险”。载于：CoRR abs/2112.04359（2021年）。arXiv：2112.04359。URL：https://arxiv.org/abs/2112.04359.[55] Jiahui Yu 等。《用于内容丰富的文本生成图像的自回归模型扩展》。2022年。arXiv：2206.10789 [cs.CV].[56] Maxim Zvyagin 等。“GenSLMs：基因组规模语言模型揭示 SARS-CoV-2 的进化动态”。载于：bioRxiv（2022年）。DOI：10.1101/2022.10.10.511571。eprint：https://www.biorxiv.org/content/early/2022/10/11/2022.10.10.511571.full.pdf。URL：https://www.biorxiv.org/content/early/2022/10/11/2022.10.10.511571。